ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 3, Issue 10, October 2016

Heterogenous Documents Using Hierarchical Dirichlet Process

[1]Gopi Krishna L, [2] Sandeep Naramgari,
[1]PH.D Scholar, Scope School, VIT University, [2] Professor, SAS School, VIT University

***Abstract:*—the hierarchical data groupings in text corpus, e.g., words, sentences, and documents, we conduct the structural learning and infer the latent themes and topics for sentences and words from a collection of documents, respectively. The relation between themes and topics under different data groupings is explored through an unsupervised procedure without limiting the number of clusters. A tree stick-breaking process is presented to draw theme proportions for different sentences. We build a hierarchical theme and topic model, which flexibly represents the heterogeneous documents using Bayesian nonparametric. Thematic sentences and topical words are extracted. In the experiments, the proposed method is evaluated to be effective to build semantic tree structure for sentences and the corresponding words. The superiority of using tree model for selection of expressive sentences for document summarization**

***Keywords:*--Bayesiannonparametrics(BNPs), latent Dirichlet allocation (LDA) tree stick-breaking process (TSBP)**

I. INTRODUCTION

Unsupervised learning has a broad goal of extracting features and discovering structure within the given data. The unsupervised learning via probabilistic topic model [1] has been successfully developed for document categorization [2], image analysis [3], text segmentation [4], speech recognition [5], information retrieval [6], document summarization [7], [8], and many other applications. Using topic model, latent semantic topics are learned from a bag of words to capture the salient aspects embedded in data collection. In this paper, we propose a new topic model to represent a bag of sentences as well as the corresponding words. As we know, the concept of topic is well understood in the community. Here, we use another related concept theme. Themes are the latent variables, which occur in different level of grouped data, e.g., sentences, and so the concepts of themes and topics are different. We model the themes and topics separately and require the estimation of them jointly. The hierarchical theme and topic model is constructed. Fig. 1 shows the diagram of hierarchical generation from documents, sentences to words given by the themes, and topics, which are drawn from their proportions. We explore a semantic tree structure of sentence-level latent variables from a bag of sentences, while the word-level latent variables are learned from a bag of grouped words allocated in individual tree nodes. We build a two-level topic model through a compound process. The process of generating words conditions on the theme assigned to the sentence. The motivation of this paper aims to go beyond the word level and upgrade the topic model by means of discovering the hierarchical relations between

The latent variables in word and sentence levels. The benefit of this model is to establish a hierarchical latent variable Model, which is feasible to characterize the heterogeneous documents with multiple levels of abstraction in different data groupings. This model is general and could be applied for document summarization and many other information systems.

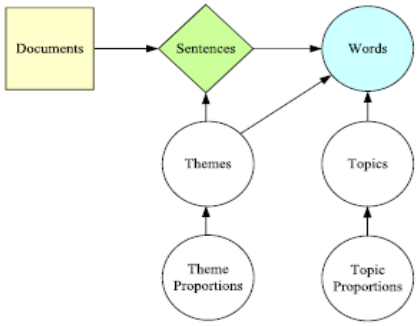

Fig. 1. Conceptual illustration for hierarchical generation of documents (yellow rectangle), sentences (green diamond), and words (blue circle) using theme and topic proportions.

II. RELATED WORK

A. Topic Model

Topic model based on latent Dirichlet allocation (LDA) [2] is constructed as a finite-dimensional mixture representation, which assumes that: 1) the number of topics was fixed and 2) different topics were independent. The hierarchical Dirichlet process (HDP) [9] and the nested Chinese restaurant process (nCRP) [10], [11] were proposed

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 3, Issue 10, October 2016

to relax these assumptions. The HDP–LDA model in [9] is a nonparametric extension of LDA, where document representation is allowed to grow structurally when more documents are observed. The number of topics is unknown a priori. Each word token within a document is drawn from a mixture model, where the hidden topics are shared across documents. DP is realized to find flexible data partitions and provide the nonparametric prior over the number of topics for each document. The base measure for the child DPs is itself drawn from a parent DP. The atoms are shared in a hierarchical way. Model selection problem is tackled through Bayesian nonparametric (BNP) learning. In the literature, the sparse topic model was constructed by decoupling the sparsity and smoothness for LDA [12] and HDP [13]. The spike and slab prior over Dirichlet distributions was applied using a Bernoulli variable for each word to indicate whether the word appears in the topic or not.

B. Topic Model Beyond Word Level

In previous methods, the unsupervised learning over text data finds the topic information from a bag of words. The mixed membership modeling is implemented for representation of words from multiple documents. The word level mixture model is built. However, unsupervised learning beyond word level is required in many information systems. For example, the topic models based on a bag of bigrams [21] and a bag of n-gram histories [5] were estimated for language modeling. In a document summarization ystem, the text representation is evaluated to select representative sentences from multiple documents. Exploring the sentence clustering and ranking [8] is essential to find the sentence-level themes and measure the relevance between sentences and documents [22]. In [23], the general sentences and specific sentences were identified for document summarization. In [7] and [24], a sentence-based topic model based on LDA was proposed to learn word-document and word-sentence associations. Furthermore, an information retrieval system retrieves the condensed information from user queries. Finding the underlying themes from documents is beneficial to organize the ranked documents and extract the relevant information. In addition to the word-level topic model, it is desirable to build the hierarchical topic model in sentence level or even in document level.

III. OBJECTIVE

In this paper, we construct a hierarchical latent variable model for structural representation of text documents. The thematic sentences and the topical words are learned from hierarchical data groupings. Each path in tree model covers from the general theme at root node to the individual themes at leaf nodes. The themes in different tree nodes contain coherent information but in varying degrees of sharing for sentence representation. We basically build a tree model for sentences according to the nCRP. The theme hierarchy is explored. The brother nodes expand the diversity of themes from different sentences within and across documents. This model does not only group the sentences into a node but also distinguish their concepts through different layers. The words of the sentences clustered in a tree node are seen as the grouped data. The grouped words in different tree nodes are driven by an HDP. The nCRP compound HDP is developed to build a hierarchical theme and topic model. To reflect the heterogeneous documents in real-world data collection, a tree stick-breaking process (TSBP) is addressed to draw austere of theme proportions. We conduct structural learning and group the sentences into a diversity of themes. The number of themes and the dependence between themes are learned from data. The words of the sentences within a node are represented by a topic model, which is drawn by a DP. All the topics from different nodes are shared under a global DP. The sentence-level themes and the word-level topics are estimated.

IV. HIERARCHICAL DIRICHLET PROCESS

HDP [9] deals with the representation of documents or grouped data where each group is associated with a mixture model. Data in different groups share a global mixture model. Each document or data grouping wd is associated with a draw from a DP given probability measure $Gd \sim DP(\alpha 0, G0)$, which determines how much a member from a shared set of mixture components contributes to that data grouping. The base measure G0 is itself drawn from a global DP by $G0 \sim DP(\gamma, H)$ with strength parameter γ and base measure H, which ensures that there is a single set of discrete components shared across data. Each DP Gd governs the generation of words $wd = \{wdi\}$ or their multinomial parameters $\{\theta di\}$ for a document. The global measure G0 and the individual measure Gd in HDP can be expressed by the mixture models with the

Shares atoms Shared atoms but different weights and given by for each d where

Shared atoms The atom φk is drawn from base measure H and the topic proportions β are drawn by SBP via $\beta|\gamma \sim GEM(\gamma)$.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 3, Issue 10, October 2016

V. HIERARCHICAL THEME AND TOPIC MODEL

Although the topic hierarchies are explored in topic model based on nCRP, only the single-level data groupings, i.e., document level, are considered in generative process. The extension of text representation to different levels of data groupings is required to improve text modeling. In addition, a single tree path cd in nCRP may not sufficiently reflect the topic variations and theme ambiguities in heterogeneous documents. A flexible topic selection is required to compensate for model uncertainty. By conducting the multiple-level unsupervised learning and flexible topic selection, we are able to upgrade system performance for document modelling. a hierarchical theme and topic model is proposed to conduct a kind of topical clustering [25], [26] over sentence level and word level, so that one can cluster sentences while clustering words. The nCRP compound HDP is presented to implement the proposed model where the text modelling in word level, sentence level and document level is jointly performed. By referring to [29], a simplified tree-structured SBP is presented to draw a sub tree td, which accommodates the theme and topic variations in document wd

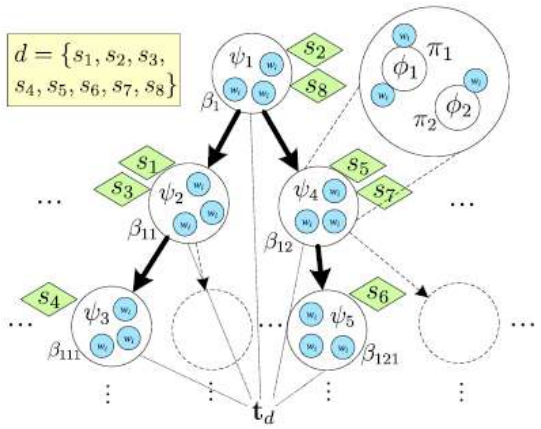

Fig. 4. Infinitely branching tree structure for representation of words, sentences, and documents based on the nCRP compound HDP

VI. CONCLUSION

This paper addressed a new hierarchical and nonparametric model for document representation and summarization. A hierarchical theme model was constructed according to a sentence-level nCRP, while the topic model was established through a word-level HDP. The nCRP compound HDP was proposed to build a tightly coupled theme and topic model, which was also seen as a theme-dependent topic mixture model. A self-organized document representation using themes in sentence level and topics in word level was developed. We presented the TSBP to draw subtree branches for possible thematic variations in heterogeneous documents. A hierarchical mixture model of themes was constructed according to the snCRP. The hierarchical clustering of sentences was implemented. The thematic sentences were allocated in tree nodes that were frequently visited. Experimental results on document modeling and summarization showed the merit of snCRP in terms of perplexity, topic coherence, and F-measure. The proposed snCRP is a general model for unsupervised structural learning. This model is generalizable to characterize the latent structure in different levels of data groupings that exist in different specialized technical data.

REFERENCES

[1] D. M. Blei, "Probabilistic topic models," Commun. ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, no. 5, pp. 993–1022, Jan. 2003.

[3] D. M. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," IEEE Signal Process. Mag., vol. 27, no. 6, pp. 55–65, Nov. 2010.

[4] J.-T. Chien and C.-H. Chueh, "Topic-based hierarchical segmentation," IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1,pp. 55–66, Jan. 2012.

[5] J.-T. Chien and C.-H. Chueh, "Dirichlet class language models for speech recognition," IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 3, pp. 482–495, Mar. 2011.

[6] J.-T. Chien and M.-S. Wu, "Adaptive Bayesian latent semantic analysis," IEEE Trans. Audio, Speech, Language Process., vol. 16, no. 1, pp. 198–207, Jan. 2008.

[7] Y.-L. Chang and J.-T. Chien, "Latent Dirichlet learning for document summarization," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Taipei, Taiwan, Apr. 2009, pp. 1689–1692.

[8] J.-T. Chien and Y.-L. Chang, "Hierarchical theme and topic model for summarization," in Proc. IEEE Int. Workshop Mach. Learn. Signal Process., Southampton, U.K., Sep. 2013, pp. 1–6.

[9] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," J. Amer. Statist. Assoc., vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

[10] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenebaum, "Hierarchical topic models and the nested Chinese restaurant process," in Advances in Neural Information Processing Systems. Vancouver, BC, Canada: MIT Press, Dec. 2004, pp. 17–24.

[11] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," J. ACM, vol. 57, no. 2, Jan. 2010, Art. ID 7.

[12] J.-T. Chien and Y.-L. Chang, "Bayesian sparse topic model," J. Signal Process. Syst., vol. 74, no. 3, pp. 375–389, Mar. 2014.

[13] C. Wang and D. M. Blei, "Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process," in Advances in Neural Information Processing Systems. Vancouver, BC, Canada: Curran & Associates Inc., Dec. 2009, pp. 1982–1989.

[14] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei, "The IBP compound Dirichlet process and its application to focused topic modeling," in Proc. 27th Int. Conf. Mach. Learn., Haifa, Israel, Jun. 2010, pp. 1151–1158.

[15] H. M. Wallach, D. M. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in Advances in Neural Information Processing Systems. Vancouver, BC, Canada: Curran & Associates Inc., Dec. 2009, pp. 1973–1981.

[16] D. I. Kim and E. B. Sudderth, "The doubly correlated nonparametric topic model," in Advances in Neural Information Processing Systems. Vancouver, BC, Canada: Curran & Associates Inc., Dec. 2011, pp. 1980–1988.

[17] A. Rodríguez, D. B. Dunson, and A. E. Gelfand, "The nested Dirichlet process," J. Amer. Statist. Assoc., vol. 103, no. 483, pp. 1131–1154, Sep. 2008.

[18] J. Paisley, L. Carin, and D. M. Blei, "Variational inference for stick-breaking beta process priors," in Proc. 28th Int. Conf. Mach. Learn., Bellevue, WA, USA, Jun. 2011, pp. 889–896.

[19] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," in Advances in Neural Information Processing Systems. Vancouver, BC, Canada: MIT Press, Dec. 2007, pp. 1353–1360.

[20] C. Wang and D. M. Blei, "Variational inference for the nested Chinese restaurant process," in Advances in Neural Information Processing Systems. Vancouver, BC, Canada: Curran & Associates Inc., Dec. 2009, pp. 1990–1998.